Research Statement

Ryan Giordano

Many researchers would be concerned if they learned that some core conclusion of their statistical analysis—such as the sign or statistical significance of some key effect—could be overturned by removing a small fraction of their data. Such non-robustness would be particularly concerning if the data were not actually drawn randomly from precisely the population of interest, or if the model may have been misspecified—circumstances that often occur in the social sciences. For example, a study of microcredit in Mexico on 16,561 households measured a negative but statistically insignificant effect of microcredit [Angelucci et al., 2015]. However, in recent work, my co-authors and I show that one can make the estimated effect of microcredit positive and statistically significant by removing just 15 households from the analysis, reversing the paper's qualitative conclusions by removing only 0.1% of the data. Since there are a combinatorially large number of ways to leave 15 datapoints out of 16,561 (over 10^{51}), finding such influential subsets by brute force would be impossible. I circumvent this difficulty by forming a *linear approximation* to the dependence of the estimator on the dataset. My technique works for a wide class of commonly used estimators, providing a fast, automatic tool for identifying small but influential subsets with finite-sample guarantees in many applied problems.

In fact, my research shows that many desirable but computationally demanding data analysis tasks are similarly amenable to automatic approximation, providing fast, easy-to-use computational tools for assessing whether small changes to modeling assumptions or datasets can meaningfully alter an analyst's substantive conclusions. As I detail below, my work encompasses Bayesian approaches (such as measuring robustness to prior specification), frequentist approaches (such as fast approximations to leave-one-out cross validation), and combinations of the two (such as measuring the robustness of Markov chain Monte Carlo procedures to data resampling). For the remainder of this statement, I will describe some of my specific past and future projects, emphasizing the ways in which I update classical results for a modern computational environment using intuitive, relevant theory, and with motivation drawn from practical applications.

1 Selected prior and ongoing work

1.1 Data sensitivity

Robustness to data ablation. In many applied settings, particularly in econometrics, a statistical analysis might be considered non-robust if it could be overturned or even reversed by removing only a small proportion of the dataset. Analyzing all possible data subsets of a certain size is typically computationally prohibitive, so I provide a method to approximately compute the number (or fraction) of observations that has the greatest influence on a given result when dropped [Broderick et al., 2020].[1] At the minimal computational cost of a single re-fit, my method provides an exact finite-sample lower bound on sensitivity to data ablation since, at worst, we have identified a sub-optimal subset to drop. I demonstrate that non-robustness to data ablation is driven by a low signal-to-noise ratio in the inference problem, is not reflected in standard errors, does not disappear asymptotically, and is not inherently a product of outliers or misspecification.

The approximation works for all M-estimators based on smooth estimating equations, a class which includes most standard estimators, including ordinary least squares, instrumental variables, generalized method of moments, variational Bayes, and maximum likelihood estimators. Using my `R` package [Giordano, 2020], the approximation is automatically computable from the specification of the estimating equation alone. By analyzing several published econometric analyses, I show that even two-parameter linear regression analyses of randomized trials can be highly sensitive. While I find some applications are robust, I show that the sign of a treatment effect can be changed by dropping less than 1% of the sample in several high-profile econometrics studies, even when standard errors are small.

[1] Following conventions in econometrics, the authors are listed alphabetically. Rachael Meager and I are equal contribution primary authors.

Approximate cross validation. The error or variability of machine learning algorithms is often assessed by repeatedly re-fitting a model with different weighted versions of the observed data; cross-validation (CV) can be thought of as a particularly popular example of this technique. In Giordano et al. [2019b], I use a linear approximation to the dependence of the fitting procedure on the weights, producing results that can be faster by an order of magnitude than repeated re-fitting. I provide explicit finite-sample error bounds for the approximation in terms of a small number of simple, verifiable assumptions. My results apply whether the weights and data are stochastic or deterministic, and so can be used as a tool for proving the accuracy of the approximation on a wide variety of problems. As a corollary, I state mild regularity conditions under which the approximation consistently estimates true leave-k-out cross-validation for any fixed k. I demonstrate the accuracy of the approximation on a range of simulated and real datasets, including an unsupervised clustering problem from genomics.

Approximately bootstrapping Bayesian posterior means. When analyzing randomly sampled data using Bayesian posteriors, it can make sense to consider the *sampling variability* of the posterior mean. For example, one might ask whether a new random sample of poll respondents in the presidential forecast model of Gelman and Heidemanns [2020] would lead to a different prediction for the election outcome. When Bayesian models are misspecified, or when they contain latent parameters on which one wishes to condition when sampling, the sampling and posterior variances can differ. In such cases, sampling variability in excess of posterior variability is symptomatic of *data non-robustness*—new data sets which are *a priori* plausible would lead to different substantive conclusions.

The sampling variability of a posterior mean can be evaluated by the bootstrap, but at the considerable cost of re-running Markov chain Monte Carlo (MCMC) hundreds of times. In Giordano and Broderick [2020, 2021], I propose an efficient alternative to bootstrapping an MCMC procedure. My approach is based on the Bayesian analogue of the influence function from the classical frequentist robustness literature. Using results from Giordano et al. [2018a, 2019b], I show that the influence function for posterior expectations can be easily computed from the posterior samples of a single MCMC procedure and state conditions under which it consistently estimates the bootstrap variance. I demonstrate the accuracy and computational benefits of my approach on an array of experiments including an election forecasting model, the Cormack–Jolly–Seber model from ecology, and a large, curated collection of models and datasets from the social sciences.

1.2 Prior sensitivity

Prior sensitivity for Markov chain Monte Carlo. Prior specification encodes key assumptions in Bayesian statistics. Often, a range of prior choices seems reasonable, and Bayesian inference can be sensitive which particular prior is used. Unfortunately, evaluating the sensitivity of Bayesian posterior expectations to prior specification by re-estimating the posterior for a large number of alternative priors is typically computationally prohibitive. In particular, Markov chain Monte Carlo (MCMC) is arguably the most commonly used computational tool to estimate Bayesian posteriors, which is made still easier by modern black-box MCMC tools such as `Stan`. However, a single run of MCMC typically remains time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models.

My software package `rstansensitivity` [Giordano, 2018, Giordano et al., 2018b], takes advantage of the automatic differentiation capacities of `Stan` together with a classical result from Bayesian robustness [Gustafson, 1996, Giordano et al., 2018a] to provide automatic hyperparameter sensitivity for generic `Stan` models from only a single MCMC run. I demonstrate the speed and utility of the package in detecting excess prior sensitivity in several examples taken from the `Stan` example datasets, including real-life datasets from the social sciences.

Prior sensitivity for discrete Bayesian nonparametrics. A central question in many probabilistic clustering problems is how many distinct clusters are present in a particular dataset and which observations cluster together. Discrete Bayesian nonparametric (BNP) mixture models address this question by placing

a generative process on cluster assignment, making the number and composition of distinct clusters amenable to Bayesian inference. However, like all Bayesian approaches, BNP requires the specification of a prior, and this prior may favor different numbers and types of posterior clusters.

In Giordano et al. [2021b], I derive and analyze prior sensitivity measures for variational Bayes (VB) approximations in general, with a practical focus on discrete BNP models. Unlike much previous work on local Bayesian sensitivity for BNP (e.g. Basu [2000]), I pay special attention to the ability of the sensitivity measures to *extrapolate* to different priors, rather than treating the sensitivity as a measure of robustness *per se*. I state conditions under which VB approximations are Fréchet differentiable functions of prior densities in a particular vector space, while also proving that VB approximations are in fact *non-differentiable* in another wide class of vector space embeddings popular in the classical Bayesian robustness literature. My co-authors and I apply the sensitivity measures to a number of real-world problems, including an unsupervised clustering problem from genomics using fastSTRUCTURE. We demonstrate that the approximation is accurate, orders of magnitude faster than re-fitting, and capable of detecting meaningful prior sensitivity in quantities of practical interest.

1.3 Improved variational inference

Uncertainty propagation in mean-field variational Bayes. Mean-field Variational Bayes (MFVB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale scientific data sets. For example, in Regier et al. [2019], my co-authors and I use MFVB to construct an approximate posterior for the identity of every astronomical object in 55TB of image data from the Sloan Digital Sky Survey. However, even when MFVB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances due to its inability to propagate Bayesian uncertainty between statistical parameters.

In Giordano et al. [2015, 2018a], I derive a simple formula for the effect of infinitesimal perturbations on MFVB posterior means, thus providing improved covariance estimates and greatly expanding the practical usefulness of MFVB posterior approximations. My approach builds on a result from the classical Bayesian robustness literature relating posterior covariances to the derivatives of posterior expectations, and includes the classical Laplace approximation as a special case. In experiments on simulated and real-life datasets, including models from ecology, the social sciences, and on a massive internet advertising dataset, I demonstrate that my method is simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude faster than MCMC.

Simplified and improved black-box variational inference. Black box variational inference (BBVI) is an easy-to-use version of variational inference requiring little more from the user than the software implementation of a differentiable log joint probability distribution. Unfortunately, standard BBVI implementations (such as the `vb` method implemented in the R software `Stan`) depend on stochastic gradient optimization, which can be more difficult to implement and assess than standard deterministic optimization methods. Further, since the exact BBVI objective function cannot be directly evaluated, one cannot compute the linear response covariances and prior sensitivity measures described above.

In Giordano et al. [2018a, 2021a] I overcome these difficulties with a simple idea: rather than use stochastic gradient to optimize the exact variational objective, one can fix a finite number of draws in advance and optimize an approximate but deterministic objective. In Giordano et al. [2021a], my co-authors and I show on a a wide range of real-life problems that the BBVI with a deterministic objective and linear response covariances produces much more accurate posterior approximations than stochastic BBVI. Further, using off-the-shelf optimization methods, we produce approximations no slower than and, often, much faster than custom implementations of stochastic BBVI. Finally, we show that the error induced by using an approximate objective is not only small but quantifiable using standard frequentist measures of uncertainty.

2 Selected future work

Though there are many potential directions for my future research, I articulate below two broad areas that I find particularly exciting.

The empirical influence function. Much of my work (particularly Giordano et al. [2019b], Broderick et al. [2020], Giordano and Broderick [2021]) has strong connections to the classical theory of von Mises expansions and the closely related concept of the influence function, which measures the effect of individual datapoints on an estimator [Mises, 1947, Reeds, 1976, Hampel, 1986, Serfling, 2009]. But my focus on the influence function evaluated at the observed data—i.e., the "empirical influence function" (EIF)—stands in contrast with much of the classical literature, which studies the asymptotic behavior of estimators via their (unobserved) limiting influence function. In our present age of automatic differentiation, large datasets, and complex models, I believe that the EIF will continue to provide practical benefits and is relatively under-studied.

In Giordano et al. [2019a], I show that higher-order EIFs can be easily and automatically evaluated and analyzed for M-estimators at a computational cost comparable to the first-order EIF—that of forming and factorizing a Hessian matrix of second-order derivatives. Thus, the EIF "amortizes" the cost of evaluating an M-estimator large number of alternative datasets: by paying a large fixed price up front (approximately computing and factorizing a Hessian matrix), one can cheaply approximate M-estimators at a very large number of alternative datasets. Natural applications of the idea include approximating the bootstrap-after-bootstrap, evaluating the sampling properties of cross-validation, and computing higher-order jackknife bias correction.

Sensitivity analysis in difficult situations. It is not always as easy to apply sensitivity analysis in practice as it is in theory. I have found that a few key problems tend to recur, and I will discuss them in turn, as well as potential solutions which draw connections to the optimization literature.

First, sensitivity analysis should, ideally, deal gracefully with incomplete optimization. For example, a collaborator from biostatistics and I have found that the popular `R` package `DESeq2` can fail in practice to fully optimize the log likelihood, and so fail to satisfy the assumptions that make sensitivity analysis possible. Instead of forcing users to optimize further, I propose that second-order empirical influence functions could simulate the effect of simultaneously taking a Newton step and perturbing the data, permitting sensitivity analysis on incompletely optimized objectives with little computation beyond that required for well-optimized objectives.

Second, the key computational bottleneck in sensitivity analysis in high-dimensional problems is typically the solution of linear systems involving the inverse Hessian of the objective function. Off-the-shelf iterative algorithms like the conjugate gradient algorithm suffice in many cases, but there is reason to believe that the present active research into stochastic second-order methods could significantly speed up sensitivity analysis in large problems.

Finally, practitioners are often interested in non-smooth objectives. For these, one promising idea is to use local approximations to speed up computationally intensive but smooth components in non-smooth problems. For example, Wilson et al. [2020] speeds up cross-validation of linear regression with a non-smooth lasso penalty by forming a fast approximation to the effect on the optimal squared error of leaving out a single datapoint, and retaining non-smoothness in the lasso penalty. We take a similar approach in Giordano et al. [2021b], retaining easy-to-compute non-linearities in posterior summary statistics.

Projects for junior collaborators. Finally, a professor must also be able to provide research topics suitable for more junior researchers, such as those who are early in their PhD. To this end, many of my existing papers can be easily "crossed" with one another, producing impactful but relatively straightforward projects for more junior collaborators. For example, I am presently working with a PhD student to combine the Bayesian influence function of Giordano and Broderick [2021] with the adversarial data ablation metrics of Broderick et al. [2020], to automatically find influential data subsets from Markov chain Monte Carlo output.

References

Angelucci, M., Karlan, D., and Zinman, J. (2015). Microcredit impacts: Evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82.

Basu, S. (2000). Bayesian robustness and Bayesian nonparametrics. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Broderick, T., Giordano, R., and Meager, R. (2020). An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*. Note: Following conventions in econometrics, the authors are listed alphabetically. Giordano and Meager are equal contribution primary authors.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the US elections. Data and model accessed Oct., 2020.

Giordano, R. (2018). StanSensitivity: Automated hyperparameter sensitivity for Stan models. GitHub repository `https://github.com/rgiordan/StanSensitivity`.

Giordano, R. (2020). Zaminfluence. GitHub repository `https://github.com/rgiordan/zaminfluence`.

Giordano, R. and Broderick, T. (2020). Effortless frequentist covariances of posterior expectations in Stan. Presentation at Stancon 2020 `https://tinyurl.com/y2e2ucp3`.

Giordano, R. and Broderick, T. (2021). The Bayesian infinitesimal jackknife for variance. *In preparation.*

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449.

Giordano, R., Broderick, T., and Jordan, M. I. (2018b). Automatic robustness measures in Stan. Presentation at Stancon 2018 `https://tinyurl.com/yyqwpowc`.

Giordano, R., Ingram, M., and Broderick, T. (2021a). Black box variational inference with a deterministic objective. *In preparation.* Giordano and Ingram are equal contribution primary authors.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2021b). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *https://arxiv.org/abs/2107.03584*. Giordano and Liu are equal contribution primary authors.

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Hampel, F. (1986). *Robust statistics: The approach based on influence functions*, volume 196. Wiley-Interscience.

Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348.

Reeds, J. (1976). *On the definition of von Mises functionals*. PhD thesis, Ph. D. Thesis, Statistics, Harvard University.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Serfling, R. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR.

Teaching Statement

Ryan Giordano

Good teaching does not only transfer knowledge, it creates intellectual community—after all, university professors are cultivating future colleagues. In an intellectual community, members are self-motivated, feedback facilitates self-improvement rather than division, and information flows freely amongst all members; analogously, good teachers contextualize course material within the students' own interests, use assessment to allow students to monitor their own progress and for the teacher to assess their own teaching, and help students teach and learn from one another rather than acting as a "sage on a stage." As I elaborate in a series of vignettes below, I find that these themes of motivation, feedback, and multi-directional communication recur and co-occur in many of the successes from my years of being a teaching assistant at the university level, as a full-time grade-school teacher in the Peace Corps, and as a volunteer math teacher with the Prison University Project.

Motivating code quality. During my second year as a PhD student at UC Berkeley, I was asked by Prof. Bin Yu to be her teaching assistant for the graduate-level course in applied statistics. Prof. Yu asked me to add a reproducible research component to the course based on my experience at Google, to which end I incorporated Github, code readability, and unit testing into the lab requirements.

I quickly realized that simply lecturing on code readability and making it a component of the grade was insufficient. The students—who were otherwise very highly motivated—simply did not see the importance of readability enough to change bad habits. To address this, I designed an in-class exercise in which the students had to "reproduce" a simple analysis written by me. In my code, I deliberately and systematically violated all the code readability guidelines I was trying to teach. As a result, it was quite difficult to understand what my analysis was doing. To sweeten the pot, I put a small but meaningful error in the code and challenged the students to find it. The students loved the puzzle-solving aspect of the assignment and, to my delight, spent much of the hour complaining about my terrible style. I had motivated the students in a way they understood, and, following this assignment, the labs' code readability improved considerably.

Evaluating in a way that accomodates multiple ability levels. Math teachers often have to accommodate a wide range of student abilities and backgrounds, and my introductory statistics class at San Quentin University through the Prison University Project (PUP) was particularly extreme in this regard. Some students had been at the top of their class when they were younger, while others were very intelligent but had only learned basic arithmetic as adults through PUP. As a consequnce, I needed to design exercises which accommodated this range of abilities and needs without leaving anyone discouraged or bored.

My solution was to reduce the proportion of the class devoted to lectures and increased the time available for individual or group work while I walked around and answered questions. I would design problem sets for such periods with the expectation that *no* student would be able to complete the whole thing in the time allotted. In this way, the faster students could quickly proceed to more challenging problems, while the slower students could spend more time with concepts that were new to them without feeling ashamed of not completing the full exercise set. When I found the same question was being asked repeatedly, I would bring everyone together for a brief lecture on the question, and then return to individual work. By providing evaluation tasks that were non-threatening and matched to students' ability levels, I helped create an inclusive classroom environment, and got a lot of feedback for myself about the effectiveness of my teaching as well.

Frequent and meaningful evaluation. When I was a teaching assistant for the graduate-level applied statistics course, the students came from a wide array of technical backgrounds, from statistics to psychology, and some students were little more than auditing, while some wanted to work hard to push their own boundaries. I expected that some students would struggle with the material, and wanted to provide evaluation that would be useful to all students.

I accomplished this goal in several ways. First, I made the rubric for grading the labs as clear as possible, and allowed for a lab to be successful in many different aspects, including clarity of exposition, quality of graphics, analytical creativity, etc. Next, I made sure that the students were continually updated with their own progress and on the grade that they were slated receive based on their performance to-date. Finally, I offered to give detailed feedback on how to improve a particular lab if the students were willing to spend time with me in person. With the help of frequent and substantive feedback, some of the students who began with the weakest backgrounds went on to become some of the strongest by the end of the class—and one such struggling student from outside statistics has even gone on to become a professional data scientist.

Short questions during lecture. Most technical lectures have many points at which minor inferential steps can be made into a short, minute-long exercise. Whenever it is possible to get feedback from the audience, I always build in such mini-exercises, which both requires students to remain actively engaged and reveals if the exposition is going too quickly. In order to allow less vocal students to participate, one can limit the number of questions per class that a given student can answer, or one can call on the third hand to be raised rather than the first.

A particularly fun variant of this idea which I developed during the Peace Corps is the "deliberate mistake." I would warn my seventh-grade math students that I was going to make a mistake in the next five minutes. The students would instantly become on the edge of their seat. In their enthusiasm, they often identified "mistakes" that were not actually mistakes, but every such instance was still a valuable teaching moment. By encouraging communication from the students to myself, I was able to both motivate the students intrinsically and evaluate for myself whether I was teaching effectively.

Two-way communication in statistical consulting. Statistical consulting, though not a classroom setting, is teaching-adjacent venue in which two-way communication is particularly important. I have provided statistical consulting services in many settings, including in the UC Berkeley statistical consulting class, as a fellow in the Berkeley Institute of Data Science, as a private contractor, and, for several years as a member of UC Berkeley's chapter of the National Security Agency Statistical Advising Group (NSASAG). Rarely, I have found, does a petitioner actually ask a useful statistical question at first, and a statistical consultant provides the most value by actively encouraging the petitioner to explain to them the problem motivation, not only its narrow statistical framing.

For example, as part of the NSASAG, we were asked how to compute low-rank approximations of matrices with some given statistical properties. Upon pressing for more information about the motivation, I learned that all that was actually needed was the computation of a t-statistic based on a linear form of a high-dimensional parameter, which I saw could be computed exactly using the conjugate gradient algorithm with no recourse to low-rank approximations. Because I promoted two-way communication rather than simply attempting to convey statistical knowledge, we were able to come to a much better solution that we would have otherwise.

Be the teacher your students want, not the teacher you would want. As a third-year undergraduate, I was asked by my engineering department to be a teaching assistant for the second-year class in statics, which was a required course for most engineering majors. Based on my own intellectual tastes at the time, I spent my weekly lectures re-deriving the course material from a more rigorous mathematical perspective in the form of assumptions and theorems. This being my first teaching role, I violated all of the above rules — I did not seek meaningful feedback from the students, I did not encourage students to work together, and, most importantly, I considered only my own motivation and not theirs. As a consequence, it was not until I received teaching evaluations at the end of the semester that I realized that the students had almost uniformly wanted more intuition from the supplementary lectures, not more rigor. Fortunately, this humbling experience set me on the long and never-ending path towards improving my teaching, leading to the more successful vignettes above.